DOCUMENT RESUME

ED 300 392 TM 012 326

AUTHOR Perkins, Kyle; Duncan, Ann
TITLE A Behavioral Anchoring Analysis of the "Iowa Tests of Basic Skills" Vocabulary and Reading Comprehension Tests.
PUB DATE 86
NOTE 34p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Education; Elementary School Students; *Grade 6; Intermediate Grades; Item Analysis; Reading Comprehension; *Reading Tests; *Standardized Tests; *Test Validity; *Vocabulary Skills
IDENTIFIERS *Anchor Tests; Behavioral Anchoring; *Iowa Tests of Basic Skills; Rasch Model

ABSTRACT

This paper reports the results of a behavioral anchoring analysis of the Iowa Tests of Basic Skills (Form G, Level 12) vocabulary (41 items) and reading comprehension (56 items) tests. Subjects were 82 sixth-grade students enrolled at the Carterville (Illinois) Grade School. More specifically, a curriculum-referenced interpretation of the ability levels is provided via definition of the characteristics of vocabulary items' stem words and, for the reading comprehension test, via definition of reading tasks and material that most students at different ability levels process successfully. The Rasch model was adopted for the analysis. For each test, anchor points on a continuum of ability levels were selected for analysis. Vocabulary items that discriminated between anchor points were batched and analyzed in terms of the stem words' frequency, distribution, abstractness, and part of speech. Reading comprehension items that discriminated between anchor points were batched and analyzed in terms of their relation to the structure of the text and the reader's prior knowledge. Results indicate a paucity of items that proved to be good discriminators. Reading should not be viewed as a listing of discrete subskills that must be independently measured. A table and four figures conclude the document. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300392

A Behavioral Anchoring Analysis of the Iowa Tests of Basic Skills Vocabulary and Reading Comprehension Tests

Kyle Perkins
Southern Illinois University

and

Ann Duncan
Carterville (IL) Grade School

(ITBS Behavioral Anchoring Analysis)

Kyle Perkins
Department of Linguistics
Southern Illinois University
Carbondale, Illinois 62901-4517

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KYLE PERKINS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM012 326

BEST COPY AVAILABLE

Abstract

This paper reports the results of a behavioral anchoring analysis of the Iowa Tests of Basic Skills (Form G, Level 12) vocabulary and reading comprehension tests. For each test, anchor points on a continuum of ability level were selected for analysis. Vocabulary items which discriminated between anchor points were batched and analyzed in terms of the stem words' frequency, distribution, abstractness, and part of speech. Reading comprehension items which discriminated between anchor points were batched and analyzed in terms of their relation to the structure of the text and the reader's prior knowledge.

INTRODUCTION

This paper presents the results of a behavioral anchoring analysis of the Iowa Tests of Basic Skills (Form G, Level 12) vocabulary and reading comprehension tests (Hieronymus, Hoover, and Lindquist, 1986). A recent development by the National Assessment of Educational Progress, behavioral anchoring was designed to provide a description of what students at different anchor points on the ability continuum could do that students at lower anchor points could not do (Messick, Beaton, and Lord, 1983; National Assessment of Educational Progress, 1985). The input for the descriptions comes from an analysis of the test items that were shown to be robust discriminators between the previously selected anchor points on the ability scale. The analysis of those test items that discriminated between what students at each point could do that students at lower levels could not do provides the necessary information for generalizing from the test items to different classes of accomplishment. In sum, behavioral anchoring has a twofold purpose: (1) to define achievements; and (2) to make test scores interpretable in terms of what students at different ability levels can and cannot do.

METHODOLOGY

Subjects and instrumentation

Eighty two sixth grade students enrolled at Carterville (Illinois) Grade School generated the data for this study. The data were collected from the Iowa Tests of Basic Skills (Form G, Level 12) battery which is administered during the spring of each school year as a part of the school's

assessment program. We do not have demographic data on these students because the item responses were coded anonymously, but it can be safely assumed that the sample of sixth graders from whom we do have data is representative of the Carterville Grade School sixth grade population.

The ITBS Form G, Level 12 vocabulary test contains 41 items; working time was fifteen minutes. According to the Preliminary Teacher's Guide (Hieronymus, Hoover, and Lindquist, 1986),

> The Vocabulary test measures both the extent or breadth of vocabulary and the ability to make fine distinctions in word meaning; that is, it measures both extensively and intensively. Two different skills classification systems reflect these two dimensions. The first is a content-area skills classification. The second is based on linguistic/structural distinctions.
>
> The content-area skills classification is an empirical, computer-generated system based on the American Heritage corpus of words sampled from texts representing a range of required and recommended reading in elementary and secondary schools.
>
> The ITBS skills classification was developed by combining the seventeen American Heritage library and curriculum categories into seven generalized categories; by entering American Heritage frequencies for all stimulus (stem) and correct response (key) words on the Vocabulary test; and by generating primary and secondary classifications for stem and key words, according to the following rules:
>
> 1. The primary classification of a word was the content-area classification under which it most frequently occurred.
> 2. A word was assigned a secondary classification if it appeared in a second content area with a frequency within ten percent of its primary classification frequency.
> 3. Words that appeared less than five times in the American Heritage list were considered unclassified.
>
> The second skills classification reflects the fact that vocabulary is a matter of concept development. The classification of skills objectives is based on functional distinctions in concept development, i.e., the part that nouns, verbs, modifiers, and connectives play in the language (p. 41).

The Form G, Level 12 Reading Comprehension test contains 56 items; working time was 42 minutes.

The Preliminary Teacher's Guide provided the following description:

> The sixteen skills objectives represented in the Reading Comprehension test are classified into three major divisions: facts-to recognize and understand stated factual details and relationships; inferences-to infer underlying relationships; generalizations-to develop generalizations from a selection. These three categories correspond closely to those of "literal meaning," "interpretative meaning," and "evaluative meaning" used by some authorities in classifying objectives of reading.
>
> Objectives involving factual details and relationships are often represented by items of the who-what-when-where types. The first three skills objectives--description, categorization, and relationships--correspond to the descriptive, categorical, and relational cognitive style modes of information processing described by Kagan, Moss, and Sigel (1963) and by Robinson and Gray (1974). The fourth objective is concerned with deducing the special meaning of words or phrases from context.
>
> The second major class of skills objectives, inferences, is represented for the most part by why? questions. The four objectives are understanding cause-and-effect relationships, drawing conclusions, inferring traits and feelings, and inferring motives and reasons for the actions of characters.
>
> Objectives in the third category, generalizations, involve higher order skills concerned with main ideas, organization, application, purpose, and viewpoint of the author, and the structure, mood, or style of the selection (pp. 44-45).

Rasch one-parameter item response theory model

To generate estimates of item difficulty and person ability for the behavioral anchoring study, we submitted the item responses from each test to the Rasch item response theory model using the PROX method (Wright and Stone, 1979). The units of measurement generated by the Rasch model are called logits which Wright and Stone described as follows:

> A person's ability in logits is their natural log odds for succeeding on items of the kind chosen to define the "zero" point on the scale. And an item's difficulty in logits is its natural log odds for eliciting failure from persons with "zero" ability (p. 17).

With the Rasch model, positive logits of person ability indicate high ability; negative logits of person ability, low ability. Positive logits of item difficulty indicate high item difficulty; negative logits of item difficulty, low item difficulty. Person ability logits are adjusted for test width expansion; item difficulty logits are adjusted for sample spread expansion. In addition, the Rasch logits are standardized so that the average person ability over all persons and the average item difficulty over all items is zero, and the standard deviation is unity. The logits are on an equal interval scale.

During the Rasch calibrations it was necessary to remove one person's responses from the vocabulary test matrix because that person scored 100 percent. Consequently, it was necessary to remove two items from the analysis because they had 100 percent responses. The final Rasch calibrations for the vocabulary test were based on 39 items and 81 persons. No items or persons were deleted from the reading comprehension test analysis for no person scored 100 percent, nor were there any items whose item scores were 100 percent.

Wright and Stone provided the rationale for deleting persons who scored 100 percent and items whose item scores were 100 percent:

> When no one in a sample of persons gets an item correct, that shows that the item is too difficult for this sample of persons. However, no further information is available as to just how much too difficult it actually is. When everyone gets an item correct, that shows that the item is too easy for these persons, but again, no further information is available as to exactly how much too easy the item actually is. To make a definite estimate for a very easy item we must find at least one measurable person who gets it incorrect, and for a very hard item, at least one measurable person who gets it correct. That is, we must "bracket" the item between persons at least one of whom is more and at least one of whom is less able than the item is difficult (p. 32).

Item response theory models are particularly well suited for behavioral anchoring research because of their invariance parameters. What this means is that the item difficulty estimates for the items are independent of the sample of students used from the population of students for whom the test was designed and intended. With regard to the estimates of person ability, the logits of person ability are independent of the particular selection of test items used from the population of items for which calibration estimates were generated.

General procedure for behavioral anchoring

For the analyses of the ITBS vocabulary and reading comprehension tests, we followed the general procedures for behavioral anchoring detailed in Beaton (1985). First, we selected anchor points for each test at which items could be found that discriminated between what students at each point on the ability continuum could do that students at lower points on the ability scale could not do. Next, we batched the items found to discriminate between pairs of anchor points. Then, we analyzed the test items and wrote a description of what the items at each level represented.

To generate some estimates for anchor points, we used the following reading proficiency (RP) scale in Beaton (1985):

$$RP = 250 + 50 \text{ (Rasch person ability logit)}$$

This scale allows us to report the results for the two tests as the estimated true score on a hypothetical test with the following idealized properties: (1) a mean proficiency of 250; (2) a standard deviation of 50; and (3)

a hypothetical range of zero to 500.

Beaton reported that for the NAEP reading scale, the following points were chosen for anchoring: 150, 200, 250, 300, and 350.

For our vocabulary responses the RP scale ranged from 181 to 411; the RP scale reading comprehension ranged from 195 to 444. We elected to use three anchoring points for each test: 200, 300, and 400. The actual RP scale anchoring points for the vocabulary test were 203 (15 correct responses, -.94 person ability logit); 297 (28 correct responses, +.94 person ability logit); and 395 (37 correct responses, +2.9 person ability logits). The actual RP scale anchoring points for the reading comprehension test were 195 (16 correct responses, -1.1 person ability logits); 302 (39 correct answers, +1.04 person ability logits); and 408 (52 correct responses, +3.16 person ability logits).

Next, for each test we selected items that discriminated between each pair of adjacent anchor points, and we used the following criteria for selecting items at each anchor point. Students at that anchor point had a probability equal to or greater than 80 percent of correctly answering the item, and students at the next lower anchor point had a probability of less than 50 percent of correctly answering the item.

To determine the response probability for each item at the three different anchor points, it was necessary to determine the differences between person ability and item difficulty and then to utilize the following table from Wright and Stone:

Probability of success p	Difference between person ability and item difficulty in logits b-d
.10	-2.2
.25	-1.1
.50	0.0
.75	1.1
.90	2.2

(p. 201).

Using these criteria an item could be selected for discriminating between only one pair of adjacent anchoring points.

Then for each test we batched the items found to discriminate between pairs of adjacent anchoring points. We then analyzed the items and wrote a description of what the items at each adjacent anchoring point represented, generalizing from each batch of items to the level of accomplishment that the items represented. We based our generalized descriptions on the following foci.

Foci for the generalized descriptions of achievement at the adjacent anchoring points on the vocabulary test

We based our generalized descriptions of vocabulary achievement on four foci: (1) a word's frequency, the number of times the word occurred in Carroll, Davies, and Richman's (1971) frequency list; (2) the word's distribution, a measure of the dispersion of word frequencies over 17 subject matter areas in the Carroll et al. list; (3) abstractness, whether the word was concrete or abstract; and (4) the word's part of speech.

The Carroll et al. frequency list was based on a corpus of 5,088,721 words drawn in 500-word samples from 1,045 texts which reflected a range of required and recommended reading to which students in grades 3 through 9 are exposed. A word's frequency is the number of times it occurred in the corpus of 5,088,721 words. The distribution value has to do with a word's being found only in a single subject matter category regardless of its frequency or its distribution being found across the 17 categories exactly proportionally to the total number of tokens (occurrences) in these categories. The 17 categories were Reading, Literature, Library Fiction, English and Grammar, Composition, Spelling, Mathematics, Library Nonfiction, Library Reference, Magazines, Social Studies, Religion, Science, Music, Art, Home Economics, and Shop.

Foci for the generalized descriptions of achievement at the adjacent anchoring points on the reading comprehension test

We wanted to base our generalized descriptions of achievement at the adjacent anchoring points for the reading comprehension test on a synthesis of recent research on some of the major constructs in reading and to use that synthesis for application in this study. Consequently, we based our descriptions on (1) the relationship of each item to the structure of the text and on (2) the reader's prior knowledge.

Text structure

We chose Meyer's (1975) prose analysis system as a guideline for a discussion of how the items related to the structure of the text. Meyer's system utilizes a detailed outline or tree structure which she termed a

content structure, and it relates all of a text's information and shows how some textual ideas are superordinate to other ideas. There are three distinct levels of content structure. The top-level structure is the rhetorical relationship that binds together all the propositions in a text and gives texture or overall organization to the text. The top third level of the total content structure also contains the main idea, the gist, the text's message, which are also call macropropositions. The middle third of the total content structure contains the major supporting details that complement the major premises of the text's message. The middle third entities are called micropropositions. The bottom third of the total content structure contains very specific minor details.

The purpose of the rhetorical relationships, according to Meyer (1984), is to "interrelate sentences, subordinate certain propositions to other propositions, and give a text its top level structure" (p. 114), and the rhetorical relationships fall into five basic groups:

> An antecedent/consequent or covariance rhetorical relationship shows a causal relationship between topics....A comparison relationship points out differences and similarities between two or more topics. Analogy, alternative, and adversative are three formats of comparison structures. A comparison:analogy relationship is exemplified in the Puppy Chow advertisement: "You should be as careful choosing a puppy food as you are choosing a baby food." A comparison:alternative relationship presents equally weighted, but mutually exclusive positions. A comparison: adversative relationship unequally emphasizes opposing viewpoints or events....The collection relationship shows how ideas or events are related together into a group on the basis of some commonality, such as a sequence of events organized by time order or space order....Description relationships give more information about a topic by presenting attributes, specifics, explanations, or settings....The response rhetorical relationship includes the remark and reply, question and answer, and problem and solution formats (pp. 114-115).

The reader's prior knowledge

Textually relevant background knowledge is a widely researched phenomenon in reading research. Farr, Carey, and Tone (1986) have noted the role which background knowledge plays in reading comprehension:

> The emerging model of comprehension asserts that comprehension is an active process in which the reader constructs meaning from text cues, calling upon knowledge of language, text structure and conventions, content concepts, and communication. This process is essentially inferential with readers using their existing knowledge to link discrete pieces of information in the text, to ascribe appropriate meanings to words, and to fill in implied information. In addition, readers monitor their state of understanding and engage various strategies according to different purposes and genre, or to encountered difficulties (p. 136).

Farr and Carey (1986) summarized the importance of prior knowledge by noting that "meaning is as dependent on the reader as it is on the text" (p. 34).

We used Schlesinger and Weiser's (1970) facet design and Pearson and Johnson"s (1978) classification scheme and operational rules for determining different comprehension categories to help us ascertain the extent to which the correct responses to the discriminating items were dependent on the reader's prior knowledge.

Schlesinger and Weiser's (1970) facet design was a test development procedure based on an a priori conceptual analysis which focuses on the explicit relationship of a reading test's items to the passages on which the items were based. Schlesinger and Weiser (197) described the facet design as follows:

> Facet design concentrates on the relationship of the item to the information presented in the text or elsewhere, rather than on skills and abilities presumably involved in comprehending the text. An items is conceived of as consisting of one or more

> statements. Each statement usually has several parts, each of which stands in the relation of agreement or contradiction with a given source of information or in some cases, the source supplies no information. The facets include form of information (explicit or implicit), source of information (the text, formal instruction, and informal experiences), context (appropriate or not appropriate) and the frequency with which information appears in the source (p. 566).

Pearson and Johnson (1978) originated a three-part classification scheme for reading comprehension items, and their classification scheme focuses on the relations between each question and its keyed response. Pearson and Johnson (1978) provided the following operational rules for determining the comprehension categories:

> A question-answer relation is classified as textually explicit if both question and answer are derivable from the text and if the relation between question and answer was explicitly cued by the language of the text (p. 163). A question-answer relation is classified as textually implicit if both question and answer are derivable from the text but there is no logical or grammatical cue tying the question to the answer and the answer given is plausible in light of the question (p. 163). Scriptally implicit (comprehension) occurs whenever a plausible nontextual response is given to a question derivable from the text (p. 164).

RESULTS

(Put Table 1 about here)

Table 1 presents the means, standard deviations, ranges, modes, medians, and estimates of reliability for the two tests. The Kuder-Richardson formula 20 was used to estimate the internal consistency reliability of the two measures. As Table 1 indicates, this sample of sixth graders averaged 27.61 correct answers on the vocabulary test, that is 67.34 percent correct. This average raw score converts to a grade-equivalency of Grade 7, third month. The average correct score on the reading comprehension test was

40.31 (71.97 percent correct), and the grade-equivalent conversion is Grade 7, seventh month. As a group, these students scored quite high; in fact, one student correctly answered all 41 vocabulary items.

(Put Figures 1 and 2 about here)

Figures 1 and 2 portray the frequency distributions for the two measures. As these figures indicate, there is some evidence of negative skew, that is, many of the student scores are clustered toward the high end of the scoring continuum. With regard to the school's ability to instill instruction as measured by the ITBS, this is not a problem, for this type of distribution is indeed desirable as an indication that many, if not a majority, of the students achieved the objectives of the instructional program.

On the other hand, skewed distributions can cause psychometric problems. A characteristic of skewed distributions is that scores tend to be compacted at one end of the scoring continuum, in this case, toward the high side. This phenomenon is referred to as boundary effects, and the result is that one has difficulty distinguishing among students in their ability at the compacted end of the scoring continuum resulting in a loss of person separability and reliability.

Vocabulary behavioral anchoring results

For the vocabulary test, the results indicated that there were no items that discriminated between anchor point 395 (37 correct responses, +2.9 person ability logits) and anchor pooint 297 (28 correct responses, +.94 person ability logit). That is, there were no vocabulary items for

which it was the case that students who scored 37 correct had a probability of 80 percent or higher of answering correctly and students who scored 28 had a probability of less than 50 percent of answering correctly, the criteria which we set in advance for robust discrimination.

There were no items which discriminated between anchor point 297 and anchor point 203 (15 correct responses, -.94 person ability logit).

Twenty two vocabulary items discriminated between anchor point 395 and 203. For these twenty two items (53.66 percent of the vocabulary test) the interpretation is that the students at anchor point 395 (37 correct responses) had a probability of 80 percent or higher of answering these items correctly and the students at anchor point 203 (15 correct responses) had less than a 50 percent probability of answering the items correctly.

In the description that follows, the interpretation is that the students at anchor point 395 achieved the abilities which are described below and the students at anchor point 203 did not. Our application and utilization of these conservative IRT parameter estimates indicates that for this particular subject pool, the ITBS vocabulary test does not have robust discriminatory power to distinguish between anchor points 395 and 297.

Vocabulary
RP scale 395
37 correct responses
+2.9 person ability logits

Students at this level of achievement could choose the correct response for four abstract adjectives which occurred on average 76 times in the Carroll et al. corpus of 5,088,721 words and were distributed in 11 of the subject matter categories with the frequencies distributed over the 11 categories exactly proportionally to the total number of tokens in those

categories; for four concrete nouns which occurred on average 288 times in the corpus and were distributed in 12 categories; for five abstract transitive verbs which occurred on average 833 times in the corpus and were distributed in 14 categories; for five abstract nouns which occurred on average 275 times in the corpus and were distributed in 13 categories; for one concrete adjective which occurred 306 times in the corpus and was distributed in 14 categories; for one abstract intransitive verb which occurred 212 times and was distributed in 14 categories; for one concrete transitive verb which occurred 12 times and was distributed in 9 categories; and for one concrete adverb which occurred 13 times and was distributed in 6 categories.

The distribution of the discriminating items by part of speech was the following: nouns, 40.91 percent; verbs, 31.82 percent; adjectives, 22.72 percent; adverbs, 4.55 percent. The distribution of the discriminating items on the abstract-concrete dichotomy was the following: abstract words, 68.18 percent; concrete words, 31.82 percent.

The students at anchor point 395 could choose the correct response for words which occurred on average 342 times in the 5,088,721-word Carroll et al. corpus and which were distributed on average in 12 of the 17 subject matter categories. The robustly discriminating vocabulary items were obviously rather low frequency words.

Reading comprehension behavioral anchoring results

Our analyses indicated that three reading comprehension items discriminated between anchor point 408 (52 correct responses, +3.16 person ability logits) and anchor point 302 (39 correct responses, +1.04 person ability logits).

Only one reading comprehension question discriminated between anchor point 302 (39 correct responses, +1.04 person ability logits) and anchor point 195 (16 correct responses, -1.1 person ability logits).

Reading Comprehension
RP scale 408
52 correct responses
+3.16 person ability logits

Students at this level of achievement could (1) develop generalizations by recognizing qualities of a text's structure with an antecedent/consequent rhetorical relationship; (2) infer a cause and effect relationship from a text with a descriptive rhetorical structure; and (3) recognize and understand stated factual details from a text with a descriptive rhetorical structure.

For these items there were no more than three occurrences of information in the texts; in fact, there was only one occurrence of information for two of the items.

All the items in this group can be characterized as manifesting textually implicit question-answer relations because there were no grammatical or logical cues tying the questions to the answers.

The sources for the correct responses for two of the three items were found in micropropositions which occurred in the middle third of a Meyer total content structure. The source for the third item was located in a macroproposition which was found in the top third of a content structure.

Reading Comprehension
RP scale 302
39 correct responses
+1.04 person ability logits

Students at this level of achievement could infer the author's purpose for writing a piece of text which listed the attributes of an object, i.e., a descriptive rhetorical organization, and the attributes were discussed in the chronological order in which the components which comprised the object were manufactured. Because the purpose of the text was to describe a common object, the student had to infer a plausible nontextual response which was derivable from the text, i.e., a scriptally implicit question-answer relation; therefore, the students had to infer a macroproposition. There were no explicit occurrences of information in the text.

DISCUSSION

One is struck by the paucity of items which proved to be good discriminators. Twenty two vocabulary items (53.66 percent of the vocabulary test) discriminated between two anchor points which were not adjacent. Three reading comprehension items (5.36 percent of the reading test) discriminated between the first and second highest anchor points, and only one item (1.79 percent of the reading test) discriminated between the second and third anchor points.

There are at least three factors which might explain the lack of discrimination: (1) boundary effects; (2) the latent trait is too restricted in its range; and (3) our potential misuse of the ITBS.

We have previously noted that the mean scores for this sample were quite high. The frequency distributions indicated that there was some evidence of skewed distributions toward the high end of the scoring continuum, a phenomenon known as boundary effects which leads to loss of person separability

and reliability.

We can paraphrase the second factor by stating that this sample's person ability range exceeded and outpaced the difficulty range of the items. We have attempted to illustrate graphically this factor in Figures 3 and 4.

(Put Figures 3 and 4 about here)

Figures 3 and 4 present definitions of the vocabulary and reading comprehension variables by item difficulty distribution and by person ability distribution in logits. The numbers in single parentheses, ( ), indicate the number of items at a particular logit of item difficulty; the numbers in double parentheses, (( )), indicate the number of persons at a particular logit of person ability.

An inspection of Figure 3 indicates the following. In the -4.3 to -3.0 logit interval there are no persons and one item; in the -2.9 to -2.0 interval, no persons and three items; in the -1.9 to -1.0 interval, one person and four items; in the -0.9 to -0.0 interval, ten persons and thirteen items; in the +0.1 to +1.0 interval, thirty five persons and eleven items; in the +1.1 to +2.0 interval, twenty seven persons and five items; in the +2.1 to +3.0 interval, five persons and one item; at the +3.2 logit there are two persons and one item. Two concepts are particularly telling with regard to the paucity of items that were robust discriminators. First, for all positive logits (difficult items and high ability persons) there were seventy persons but only eighteen items. At logit +0.2, the mode of the distribution, there were no items which suggests to us that for this particular sample of sixth graders, the ITBS vocabulary test is weakest

at the mode of our sample.

As Figure 4 indicates, we have similar problems with the reading comprehension test. In the -3.7 to -3.0 interval, there were no persons and one item; in the -2.9 to -2.0 interval, no persons and one item; in the -1.9 to -1.0 interval, one person and ten items; in the -0.9 to -0.0 interval, seven persons and fourteen items; in the +0.1 to +0.9 interval, twenty five persons and seventeen items; in the +1.0 to +1.9 interval, twenty two persons and twelve items; in the +2.0 to +2.9 interval, twenty persons and one item; and in the +3.0 to +3.9 interval, seven persons and no items. A most vexing issue is that from +1.0 to +3.9 logits there are forty nine persons but only thirteen items. It becomes quite clear that, if we want to discriminate among the majority of persons found in the +1.0 to +3.9 range, then we must have additional items which have item difficulties appropriate to the ability levels of these persons. For this particular piece of research, given the high person ability range of the students, perhaps we should have administered Level 13 to improve the discriminability range of the test.

The previous discussion naturally leads to our third factor which may explain the paucity of items which were robust discriminators. Perhaps we have expected too much from the ITBS in that we were expecting several items to discriminate between the various anchor points. The ITBS is a standardized achievement test, and such instruments are designed to provide indices of what students know at a particular point in time, which, in fact, the ITBS does quite admirably. The point here is that a single achievement battery cannot be best suited for all purposes, e.g., discriminating between different anchor points based on IRT parameters.

SUMMARY

In conclusion, we have attempted to provide a curriculum-referenced interpretation of ability levels by defining the characteristics of the vocabulary items' stem words and for the reading comprehension test, by defining the reading tasks and the reading material that a majority of students at the different ability levels had processed successfully. In deriving the reading comprehension behavioral anchoring descriptions we kept in mind the proposition that reading comprehension is a dynamic, interactive process which integrates "a variety of declarative, procedural, aand strategic knowledge in different ways, depending on the state of the comprehending system and the information available to it from various sources" (Johnston, 1984, p. 160). In other words, reading should not be viewed as a listing of discrete, separate subskills which must be measured independently. For these reasons, we elected to focus our descriptions around the items' keyed responses' relationship to the structure of the text and the reader's prior knowledge.

PEDAGOGICAL NOTE

Although our focus in this paper was not pedagogical, the results do have pedagogical implications. Our Rasch analysis of the reading comprehension test could be of immense benefit to the seventh grade teachers who instruct these students. The Rasch person ability estimates could be used to group these students for instructional purposes, in particular, for individualized instruction. Such individualization is becoming more of a possibility because standardized achievement tests are being matched

to textbooks. If the seventh grade materials were calibrated, then the curriculum could be fitted to the students rather than fitting the students to the curriculum.

REFERENCES

Beaton, A. Implementing the New Design: The NAEP Technical Report. Report No. 15-TR-20. Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service, 1985.

Carroll, J. B., Davies, P., and Richman, B. Word Frequency Book. Boston: Houghton Mifflin, 1971.

Farr, R., and Carey, R. F. Reading: What Can Be Measured?, second edition. Newark, DE: International Reading Association, 1986.

Farr, R., Carey, R. F., and Tone, B. "Recent Theory and Research into the Reading Process: Implications for Reading Assessment." In Reading Comprehension: From Research to Practice, edited by J. Orasanu, pp. 135-149. Hillsdale, NJ: Lawrence Erlbaum Associates, 1986.

Hieronymus, A. N., Hoover, H. D., and Lindquist, E. F. Iowa Tests of Basic Skills (Form G, Level 12). Chicago: The Riverside Publishing Company, 1986.

Hieronymus, A. N., Hoover, H. D., and Lindquist, E. F. Preliminary Teacher's Guide. Chicago: The Riverside Publishing Company, 1986.

Johnston, P. "Assessment in Reading." In Handbook of Reading Research, edited by P. D. Pearson, pp. 147-182. New York and London: Longman, 1984.

Kagan, J., Moss, H. A., and Sigel, I. E. "Psychological Significance of Styles of Conceptualization." Society for Research in Child Development Monographs, vol. 28, no. 2 (1963), pp. 73-112.

Messick, S., Beaton, A., and Lord, F. National Assessment of Educational Progress Reconsidered: A New Design for a New Era. NAEP Report 83-1. Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service, 1983.

Meyer, B. J. F. The Organization of Prose and its Effect on Memory. Amsterdam: North-Holland Publishing Company, 1975.

Meyer, B. J. F. "Organizational Aspects of Text: Effects on Reading Comprehension and Applications for the Classroom." In Promoting Reading Comprehension, edited by J. Flood, pp. 113-137. Newark, DE: International Reading Association, 1984.

National Assessment of Educational Progress. The Reading Report Card. ETS Report No. 15-R-01. Princeton, NJ: Author, Educational Testing Service, 1985.

Pearson, P. D., and Johnson, D. D. Teaching Reading Comprehension. New York: Holt, Rinehart and Winston, 1978.

Robinson, J. E., and Gray, J. L. "Cognitive Style as a Variable in School Learning." Journal of Educational Psychology, vol. 66, no. 5 (1974), pp. 793-799.

Schlesinger, I. M., and Weiser, Z. "A Facet Design for Tests of Reading Comprehension." Reading Research Quarterly, vol. 5, no. 4 (1970), pp. 556-580.

Wright, B. D., and Stone, M. H. Best Test Design: Rasch Measurement. Chicago: MESA Press, 1979.

Table 1
Means, standard deviations, and estimates of reliability

Test	No. of subjects	No. of items	Mean	Standard deviation	Range	Mode	Median	Estimate of reliability
Vocabulary	82	41	27.61	6.00	12 to 41	23	28	.824
Reading comp.	82	56	40.31	8.73	16 to 54	50 and 52	41	.888

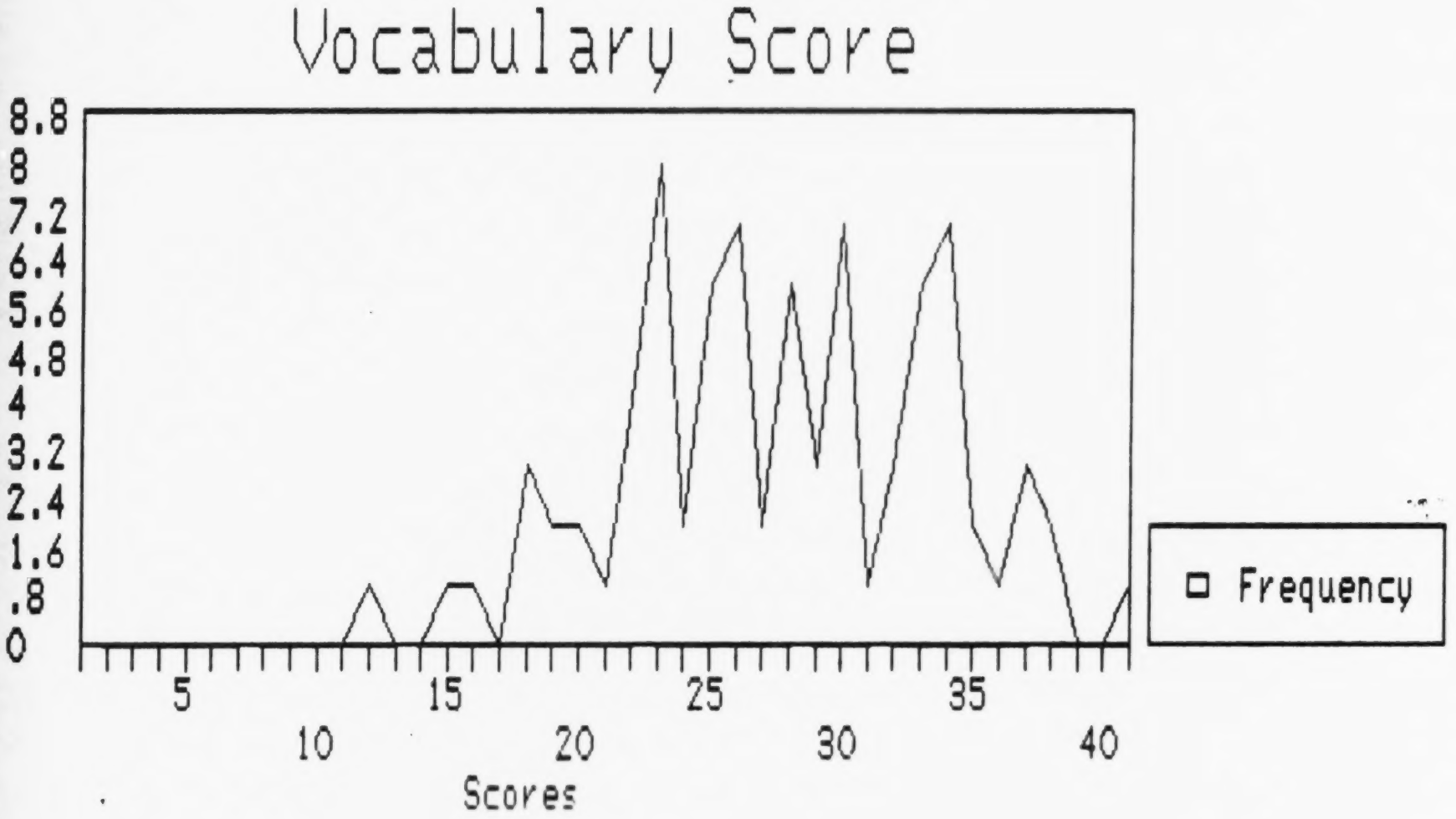

Figure 1: Vocabulary test score distribution

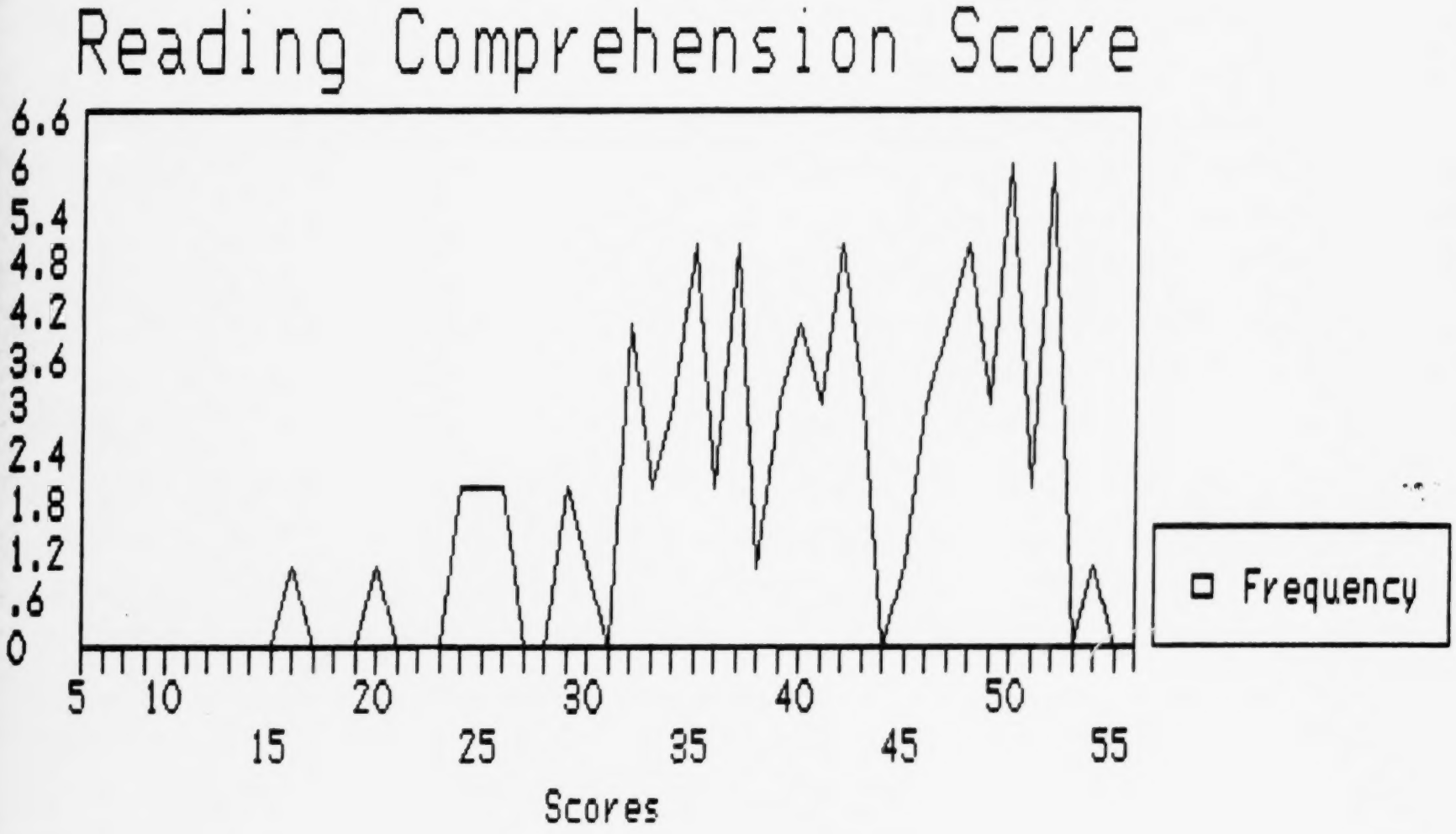

Figure 2: Reading comprehension test score distribution

Figure 3
Defining the vocabulary variable by item difficulty distribution and by person abili ty distribution
(number of items)
((number of persons))

-4.3 (1)	-1.8	+0.7 (1) ((7))	+3.2 (1) ((2))
-4.2	-1.7	+0.8 ((2))	
-4.1	-1.6 (1)	+0.9 ((6))	
-4.0	-1.5	+1.0 (2)	
-3.9	-1.4 (2) ((1))	+1.1 ((3))	
-3.8	-1.3 (1)	+1.2 ((7))	
-3.7	-1.2	+1.3	
-3.6	-1.1	+1.4 ((1))	
-3.5	-1.0	+1.5 (1)	
-3.4	-0.9 (1) ((1))	+1.6 ((3))	
-3.3	-0.8 (1) ((1))	+1.7 (1)	
-3.2	-0.7	+1.8 (1) ((6))	
-3.1	-0.6 (3)	+1.9	
-3.0	-0.5 ((3))	+2.0 (2) ((7))	
-2.9	-0.4 (1)	+2.1	
-2.8	-0.3 ((2))	+2.2	
-2.7	-0.2 (2) ((2))	+2.3 ((2))	
-2.6	-0.1 (5) ((1))	+2.4	
-2.5	-0.0	+2.5 ((1))	
-2.4 (2)	+0.1 ((4))	+2.6	
-2.3	+0.2 ((8))	+2.7 (1)	
-2.2	+0.3 (1) ((2))	+2.8	
-2.1	+0.4 (4)	+2.9 ((3))	
-2.0 (1)	+0.5 (1) ((6))	+3.0	
-1.9	+0.6 (2)	+3.1	

Figure 4
Defining the reading comprehension variable by item difficulty and by person ability distribution
(number of items)
((number of persons))

-3.7 (1)	-1.3	+1.1 (3) ((4))	+3.5
-3.6	-1.2 (3)	+1.2 ((3))	+3.6
-3.5	-1.1 (1) ((1))	+1.3 (6) ((5))	+3.7
-3.4	-1.0 (1)	+1.4	+3.8
-3.3	-0.9 (3)	+1.5 ((3))	+3.9 ((1))
-3.2	-0.8	+1.6 (1)	
-3.1	-0.7 ((1))	+1.7 (1) ((1))	
-3.0	-0.6 (2)	+1.8	
-2.9	-0.5 (3)	+1.9 ((3))	
-2.8	-0.4 (1)	+2.0 (1) ((4))	
-2.7	-0.3 ((2))	+2.1	
-2.6	-0.2 (2) ((4))	+2.2 ((5))	
-2.5	-0.1	+2.3	
-2.4	-0.0 (3)	+2.4 ((3))	
-2.3	+0.1 (1) ((2))	+2.5	
-2.2 (1)	+0.2 ((1))	+2.6 ((6))	
-2.1	+0.3 (3) ((4))	+2.7	
-2.0	+0.4 ((2))	+2.8 ((2))	
-1.9 (3)	+0.5 (4)	+2.9	
-1.8	+0.6 (4) ((3))	+3.0	
-1.7 (2)	+0.7 (1) ((7))	+3.1	
-1.6	+0.8 (2) ((5))	+3.2 ((6))	
-1.5	+0.9 (2) ((1))	+3.3	
-1.4	+1.0 (1) ((3))	+3.4	